REVISITING PRIORITIES: IMPROVING MIR EVALUATION PRACTICES

Bob L. Sturm
Centre for Digital Music, Queen Mary University of London

ABSTRACT

While there is a consensus that evaluation practices in music informatics (MIR) must be improved, there is no consensus about what should be prioritised in order to do so. Priorities include: 1) improving data; 2) improving figures of merit; 3) employing formal statistical testing; 4) employing cross-validation; and/or 5) implementing transparent, central and immediate evaluation. In this position paper, I argue how these priorities treat only the symptoms of the problem and not its cause: MIR lacks a formal evaluation framework relevant to its aims. I argue that the principal priority is to adapt and integrate the formal design of experiments (DOE) into the MIR research pipeline. Since the aim of DOE is to help one produce the most reliable evidence at the least cost, it stands to reason that DOE will make a significant contribution to MIR. Accomplishing this, however, will not be easy, and will require far more effort than is currently being devoted to it.

1. CONSENSUS: WE NEED BETTER PRACTICES

I recall the aims of MIR research in Sec. 1.1, and the importance of evaluation to this pursuit. With respect to these, I describe the aims and shortcomings of MIREX in Sec. 1.2. These motivate the seven evaluation challenges of the MIR "Roadmap" [23], summarised in Sec. 1.3. In Sec. 1.4, I review a specific kind of task that is representative of a major portion of MIR research, and in Sec. 1.5 I look at one implementation of it. This leads to testing a causal model in Sec. 1.6, and the risk of committing the sharpshooter fallacy, which I describe in Sec. 1.7.

1.1 The aims of MIR research

MIR research aims to connect real-world users with music and information about music, and to help users make music and information about music [3]. One cannot overstate the importance of *relevant* and *reliable* evaluation to this pursuit: [1] we depend upon it to measure the effectiveness of our algorithms and systems, compare them against the state of the art, chart the progress of our discipline, and discriminate promising directions from dead ends. The design of any machine listening system [2] involves a series of complex decisions, and so we seek the best evidence to guide this process. This motivates the largest contribution so far to evaluation methodology in MIR research: the Music Information Retrieval Exchange (MIREX) [9].

1.2 MIREX

MIREX represents a significant advance beyond the inconsistency of evaluation practices in the early years of MIR. Its guiding precepts include [8]: test collections should be of considerable size and private, if possible; evaluations should be performed by a private centralised system; formal statistical testing should be used to detect significant differences between submissions; results should be publicly archived; and MIREX is not a competition but an opportunity to exchange knowledge. MIREX is now a decade old (evaluating nearly 3000 submissions so far) and is linked to a significant amount of research [5]. However, MIREX suffers serious problems [11, 17, 20, 22, 23, 28, 34, 35]: its tasks can lack consideration of the user; its tasks can be poorly defined and contrived; its metrics can lack relevance; and its evaluations can lack validity.

1.3 The "Roadmap" for MIR [23]

MIREX has certainly helped MIR advance, but its evaluation practices must be improved. This fact is officially acknowledged in the 2013 "Roadmap for Music Information Research" [23], authored by 17 recognised MIR researchers at seven major institutions. Section 2.6 of the Roadmap identifies seven specific challenges related to evaluation that the discipline should address to ensure its continued development. We need to:

$\mathbf{R}_{\mathrm{I}}$ "Define meaningful evaluation tasks"
$\mathbf{R}_{\mathrm{II}}$ "Define meaningful evaluation methodologies"
$\mathbf{R}_{\mathrm{III}}$ "Evaluate whole MIR systems"
$\mathbf{R}_{\mathrm{IV}}$ "Promote evaluation tasks using multimodal data"
$\mathbf{R}_{\mathrm{V}}$ "Promote best practice evaluation methodology"
$\mathbf{R}_{\mathrm{VI}}$ "Implement sustainable MIR evaluation initiatives"
$\mathbf{R}_{\mathrm{VII}}$ "[Promote reproducible] MIR."

I identify R_{I} and R_{II} as the "linchpins." It is thus essential to define, "define" and "meaningful" for both "tasks" and "methodologies". For R_{I}, the Roadmap suggests a task is "meaningful" when it is "relevant" to a well-defined

[1] An "evaluation" is a protocol for testing a hypothesis or estimating a quantity. An evaluation is "relevant" if it logically addresses the investigated hypothesis or quantity. An evaluation is "reliable" if it results in a repeatable and statistically sound conclusion.

[2] A machine listening system is a fixed map from a recording universe to a semantic universe [30]. See Sec. 3.1.

user community, and defined (or addressed) [3] "according to some agreed criteria." For R_{II}, the Roadmap suggests an evaluation methodology is "meaningful" if it creates knowledge leading to the improvement of MIR systems, and the discipline as a whole.

1.4 The "Audio Classification (Train/Test)" task

MIREX shows MIR is replete with tasks, but I will focus on one kind: "Audio Classification (Train/Test)." This task involves building systems using feature extraction algorithms, supervised machine learning algorithms, and a training dataset, and then testing with a testing dataset. [4] At its most base, the goal is to build a system that reproduces the most ground truth of a testing dataset. This task appears in over 400 publications addressing "music genre recognition" [26, 27] (not to mention work addressing "music similarity," "music mood recognition" and "autotagging" [24]), and so typifies a major portion of MIR research. Referring to the aims of MIR and the Roadmap, I ask: how does reproducing dataset ground truth provide relevant and reliable knowledge about a system, and how to improve it, for a well-defined user-community?

1.5 Three systems designed for a specific problem

Consider three systems expressly designed to address the problem intended by the BALLROOM dataset [7]: to extract and learn "repetitive rhythmic patterns" (RRPs) from recorded audio. BALLROOM has 698, 30-second monophonic music excerpts downloaded in 2004 from a commercial web resource devoted to ballroom dancing. A system solving this problem maps 30-s music recordings to several classes, e.g., Cha cha, *according to RRPs*.

Three systems trained and tested with the same BALLROOM partitioning produce the following figures of merit: system A ($\mathscr{S}_A$) reproduces 93.6% of the ground truth; $\mathscr{S}_B$, 91.4%; and $\mathscr{S}_C$, only 28.5%. Given that a random selection of each class will reproduce about 14.3% of the ground truth, two possible hypotheses are:

H1 $\mathscr{S}_A$, $\mathscr{S}_B$ are identifying RRPs in BALLROOM, and $\mathscr{S}_C$ is not identifying RRPs in BALLROOM

H2 The features used by $\mathscr{S}_A$, $\mathscr{S}_B$ are powerful for identifying RRPs in BALLROOM, but not those of $\mathscr{S}_C$.

Let us look under the hood of each system, so to speak. Each is composed of a feature extraction algorithm (mapping the audio sample domain to a feature space), and a classification algorithm (mapping the feature domain to the semantic (label) space) [30]. For $\mathscr{S}_A$, the feature is global tempo (possibly with an octave error), and the classifier is single nearest neighbour in the training dataset [29]. For $\mathscr{S}_B$, the feature is the 800-dimensional lagged autocorrelation of an energy envelope, and the classifier is a deep neural network [18]. For $\mathscr{S}_C$, the feature is *umpapa presence*, [5] and the classifier is a decision tree.

[3] This is ambiguous in the Roadmap.

[4] A dataset is a sequence of observation, label pairs, $((r_i, s_i))$, where r_i is the ith observation and s_i is its ground truth.

[5] A fictional quantitative measure of the RRP, "OOM-pah-pah."

There are a few startling things. First, while $\mathscr{S}_A$ reproduces the most ground truth, it does so by using *only* tempo. Since tempo is not rhythm, $\mathscr{S}_A$ does not address the problem for which it was designed. [6] Second, $\mathscr{S}_C$ reproduces the least ground truth of the three, but is using a feature that is relevant to the problem intended by BALLROOM [7]. Its accuracy is so low because only the Waltz-labeled recordings have high umpapa presence, while all the others are in common or duple meter and have low umpapa presence. It seems then that we must doubt H1 and H2 when it comes to $\mathscr{S}_A$ and $\mathscr{S}_C$. What about $\mathscr{S}_B$?

$\mathscr{S}_B$ is using a feature that should contain information closely related to RRPs: periodicities of acoustic stresses observed over 10 second periods [18]. In fact, it is easy to visually interpret these features in terms of tempo and meter. The deep neural network in $\mathscr{S}_B$ is not so easy to interpret. However, given the impressiveness of recent results from the deep learning revolution [15], it might seem reasonable to believe $\mathscr{S}_B$ reproduces so much ground truth because it has learned to identify the high-level RRPs that characterise the rhythms of BALLROOM labels.

1.6 An intervention into a causal model

Let us claim that $\mathscr{S}_B$ reproduces BALLROOM ground truth by detecting RRP in the music recordings. We thus propose the causal model shown in Fig. 1(a). Consider an experiment in which we perform an intervention at the exogenous factor. We take each BALLROOM testing recording and find the minimum amount of pitch-preserving time-stretching [7] (thereby producing a tempo change only) for which $\mathscr{S}_B$ produces an incorrect class.

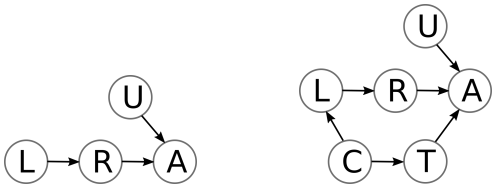

Figure 1. Two causal models relating factors: BALLROOM label (L), RRP (R), audio observation (A), competition rules (C), tempo (T), and exogenous (U).

We find [29, 31] that a mean tempo increase of 3.7 beats per minute (BPM) makes $\mathscr{S}_B$ classify all Cha-cha-labeled testing recordings as "tango." While $\mathscr{S}_B$ initially shows a very good "rumba" F-score (0.81), it no longer identifies the Rumba-labeled recordings when time-stretched by at most $\pm 3\%$. By submitting all testing recordings to a tempo change of at most $\pm 6\%$, $\mathscr{S}_B$ goes from reproducing 91.4% of the ground truth to reproducing as little as random guessing (14.3%). (By the same transformation, we can also make $\mathscr{S}_B$ reproduce all ground truth.) What is more, $\mathscr{S}_B$ assigns different labels to the same music when we change only its tempo: it classifies the same Cha cha-labeled music as "cha cha" when its tempo is 124 BPM, then "quickstep" when its tempo is 108 BPM, and so on.

[6] In fact, Dixon et al. [7] design their systems to be tempo-invariant.

[7] We use `http://breakfastquay.com/rubberband`

These results thus cast serious doubt on H1 and H2 for $\mathscr{S}_B$. It seems that a belief in H1 and H2 is not so reasonable after all: $\mathscr{S}_A$ and $\mathscr{S}_B$ reproduce BALLROOM ground truth using a characteristic that is not rhythm. It is only $\mathscr{S}_C$ that is addressing the problem intended by BALLROOM, but it does not reproduce a large amount of ground truth simply because it is sensitive to only one kind of RRP.

1.7 On the sharpshooter fallacy

A typical response to the above is that if a system can reproduce ground truth by looking at tempo, then it should. In fact, the 2014 World Sport Dance Federation rules [8] provide strict tempo ranges for competitions featuring the dances represented by BALLROOM; and the tempi of the music in BALLROOM adhere to these ranges (with the exception of Rumba and Jive) [29, 31]. This response, however, commits the *sharpshooter fallacy*: it moves the bullseye post hoc, e.g., from "extract and learn RRPs from recorded audio" to "reproduce the ground truth." [9]

That there exists strict tempo regulations for dance competitions, and that the origin of BALLROOM comes from a commercial website selling music CDs for dance competitions, motivate the alternative causal model in Fig. 1(b). This model now shows a path from the music heard in a BALLROOM recording to its ground truth label via competition rules, which explains how $\mathscr{S}_A$ and $\mathscr{S}_B$ reproduce BALLROOM ground truth without even addressing the intended problem.

1.8 Intermediate conclusion

There are of course limitations to the above. BALLROOM is one dataset of many, and in fact could be used for a different problem than RRP. MIR tasks are broader than "Audio Classification (Train/Test)," and involve many other kinds of information than rhythm. Classification accuracy is just one measure; a confusion table could provide a more fair comparison of the three systems. I use BALLROOM and the three systems above simply because they clearly demonstrate problems that can arise even when a task and problem appear to be well-defined, and a dataset is cleanly labeled and has reputable origins. Though several systems are trained and tested in the same dataset with the express purpose of solving the same problem (as is the case for all MIREX tasks of this kind), they in fact may be solving different problems. Seeking the cause of a system's performance, e.g., through intervention experiments [25, 26, 29, 31], can then reveal a confounding of "reproduce ground truth" with, e.g., "learn to recognise rhythm." It is tempting to then move the bullseye, but doing so weakens one's contribution to the aims of MIR research. Emphasising ground truth reproduction over solving intended problems can lead to promoting non-solutions over solutions with respect to the aims of MIR research. Clearly then, MIR evaluation practices must be improved, but what should be prioritised to do so?

[8] `https://www.worlddancesport.org/Rule/Competition/General`

[9] Recall these three systems were expressly designed to address the problem intended by BALLROOM (Sec. 1.5, and described in [7]).

2. NON-CONSENSUS: OUR PRIORITIES

While problems with MIR evaluation have been known for some time, there is no consensus on what should be prioritised to solve them. I now discuss several of these.

2.1 We need to collect more data

Perhaps the most immediate answer to evaluation problems is to increase the sizes of datasets. The underlying belief is that experimental power increases with the number of observations. Along with the pursuit of model generalisation, this has motivated the creation of the Million Song Dataset [2] and AcousticBrainz [19]. The advent of crowdsourcing makes data collection seem cheap, but the actual costs can be very high. First, music recordings have many layers of intellectual property, which limit their use and distribution. This directly opposes research that is open and reproducible, and so imposes a high cost to progress. Second, and most importantly, making data bigger does not necessarily improve an experiment's power, but it *certainly* increases its cost. Even if BALLROOM had 1 billion music recordings meeting competition tempo regulations, the conclusions of Secs. 1.5-1.6 would not change. The most important question to ask then is not how to collect the most data, but how to collect and use data such that it results in the most relevant and reliable evidence possible while minimising the cost incurred.

2.2 We need to find better figures of merit

A highly discussed topic of evaluation is that of metric or measure (figure of merit, FoM). Which FoM (accuracy, F-score, AUC, etc.) gives the best indication of how well a system is addressing the intended problem? MIREX typically reports several FoM in each of its tasks since a diversity of viewpoints can inform interpretation. Still, one particular FoM can dominate a research problem, e.g., classification accuracy is the most-used FoM reported in "music genre recognition" research (appearing in over 80% of such publications [27]). This is troubling since Secs. 1.5-1.6 show that the amount of ground truth reproduced by a system could say nothing relevant or reliable about its success for some problem thought well-posed. The choice of FoM is certainly important, but the most important question to ask before selecting an FoM is how to measure its relevance and reliability, and how to compare it in meaningful ways, with respect to the intended problem. Recent work is addressing this important question, e.g., [6, 11, 16].

2.3 We need to perform more formal statistical testing

Some have argued that MIR research should adopt rigorous statistical procedures [10,34]. Null hypothesis significance testing provides a remarkable set of useful tools, despite the problems that come with their interpretation [4,14]. For instance, the probability that $\mathscr{S}_B$ in Sec. 1.5 reproduces its 91.4% of BALLROOM ground truth given it is actually selecting randomly is $p < 10^{-138}$ (by a binomial test). Though one may safely reject this hypothesis, it still gives no reason to claim $\mathscr{S}_B$ is identifying RRP. Sec. 1.6 shows $\mathscr{S}_B$ to be exploiting a third way to reproduce BALLROOM

ground truth. No statistical test will improve the relevance and reliability of the measurements of the three systems in Sec. 1.5 with respect to the hypotheses posed. A different kind of evaluation must be employed. While statistical testing is useful, the most important questions to ask first are: 1) whether the evidence produced by an evaluation will be relevant and reliable at all; and 2) what statistical test is relevant to and permitted by an experiment [1, 12].

2.4 We need to use more cross-validation

In the directions of using data in smarter ways and statistical hypothesis testing, cross-validation (CV) is a testing protocol that is standard in machine learning [13]. CV holds constant a learning algorithm but changes training and testing datasets multiple times, both of which are culled from a larger dataset. Many variants exist, but the main motivations are the same: to simulate having more data than one actually has, and to avoid overfitting models. CV produces point estimates of the expected generalisation of a learning algorithm [13], but its relevance and reliability for gauging the success of a given system for solving an intended problem can be unclear. Returning to the three systems in Sec. 1.5, CV in BALLROOM will produce more estimates of the expected generalisation of the learning algorithms, but that does not then make the results relevant to the hypotheses posed. The most important question to ask then is how to design a testing protocol that can produce relevant and reliable evidence for the hypotheses under investigation.

2.5 We need to develop central, transparent and immediate evaluation

The nature of a computerised research discipline is such that one can train and test thousands of system variants on millions of observations and produce quantitative results within a short time scale. MIREX provides one vehicle, but it occurs only once a year, and has problems of its own (Sec. 1.2). This motivates commendable efforts, such as the Networked Environment for Music Analysis [36], and `mir_eval` [20]. Their aim is to increase transparency, standardise evaluation, reduce delay, mitigate legal jeopardy, and facilitate progress in MIR. Concerns of the transparency and immediacy of an evaluation, however, are premature before designing it to be as relevant and reliable as possible at the least cost.

2.6 Intermediate conclusion

The overriding priority is not to collect more data, but to develop ways to collect and use data in provably better ways. The overriding priority is not to find better FoM, but to develop ways to judge the relevance of any FoM, and to make meaningful comparisons with it. The overriding priority is not to perform more formal statistical testing, to use CV, or facilitate transparency and immediacy in evaluation, but to develop ways of producing relevant evidence while satisfying requirements of reliability and cost. This leads me to propose a principal priority for improving evaluation practices in MIR.

3. THE PRINCIPAL PRIORITY IN EVALUATION

The *principal priority* is to develop a formal framework of evaluation that facilitates a meaningful evaluation methodology for any problem that will result in relevant and reliable evidence of the effectiveness of our algorithms and systems, facilitate comparisons with the state of the art, chart the progress of the discipline, and discriminate promising directions from dead ends, all with respect to the aims of MIR research. I propose that this can be accomplished by leveraging the established design and analysis of experiments (DOE) in tandem with an effort to reign in the ambiguity of MIR problems and tasks. This will then address the shortcomings of MIREX (Sec. 1.2), position MIR to meet the Roadmap challenges (Sec. 1.3), and enable a new and progressive research pipeline.

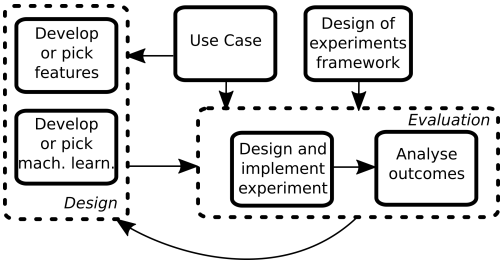

Figure 2. This new research pipeline involves a use case, a DOE framework, and feedback to improve system design.

Figure 2 illustrates a new way to engineer MIR systems and their component technologies. Three central components are the use case, a formal design of experiments (DOE) framework, and feedback from the evaluation to system design. A *use case* is a formal expression of the problem a system is to address. The *DOE framework* provides the theoretical underpinning for designing, implementing and analysing the most relevant and reliable evaluation of a system with respect to a use case at the least possible cost. These two components feed into realising an evaluation of a specific MIR system, itself built with reference to the use case. The evidence produced by the evaluation thus leads to improving the design of the MIR system. The use case together with the DOE framework temper linchpins R_I and R_{II}. With these firmly established, the rest of the Roadmap evaluation challenges can be accomplished.

3.1 On the use case

I define a use case in [30] as a means for mitigating ambiguity in research, and thus tempering linchpin R_I. For instance, nearly all published work on the problem of "music genre recognition" does not explicitly define the problem, instead posing it as reproducing the ground truth of a given dataset. As a result, many hundreds of publications have unknown relevance to the aims of MIR (Sec. 1.1), even when they use the same dataset [26].

A *use case* is defined as a tuple of four formal elements: the music universe (Ω), the music recording universe ($\mathcal{R}_\Omega$), the description universe ($\mathcal{S}_{\mathcal{V},A}$), and a set of success criteria. This retains a distinction between the intangible Ω and

ID	*Treatments* ($\mathcal{T}$)	*Experimental unit*	*Observational unit* ($\omega \in \Omega$)	*Treatment structure*	*Plot structure*	*Response*	*Response model*
a	Amounts of compost & water	tomato plant in a greenhouse pot	tomato plant in a greenhouse pot	all combinations of two factors	blocks	tomato yield (grams)	simple textbook
b	New feed, old feed	pen	calf	new treatment and control	unstructured	weight (kg)	simple textbook
c	Local or remote learning	students in DOE 101 classroom-year	student	unstructured	blocks	test score (%)	fixed effects
d	Four wines	judge	judge-tasting	unstructured	unstructured	score $\{1, ..., 5\}$	simple textbook

Table 1. Examples of the various components of experiments. (a): estimating the relationship between tomato yield, and the amount of water and compost applied to a tomato plant in a greenhouse pot. (b): testing for a significant difference between new and old feed in the weight gain of a calf (several calves to a pen, feed applied to whole pen). (c): testing for a significant difference between students learning DOE locally or remotely (students are or are not math majors, thus defining two blocks, motivating a fixed effects response model). (d): testing for a significant difference in wine quality.

the tangible $\mathcal{R}_\Omega$ — elements of which are fed to a recorded music description system (a map, $\mathscr{S} : \mathcal{R}_\Omega \to \mathcal{S}_{\mathcal{V},A}$). $\mathcal{S}_{\mathcal{V},A}$ is a set of elements assembled in a meaningful way for a user. The success criteria embody the requirements of a user for mapping from Ω and/or $\mathcal{R}_\Omega$ to $\mathcal{S}_{\mathcal{V},A}$.

To provide illustration, let us define two use cases. Define Ω as all music meeting specific tempo and stylistic regulations with respect to the labels in BALLROOM. Define $\mathcal{R}_\Omega$ as the set of all possible 30-s, monophonic recording excerpts of the music in Ω sampled at 22050 Hz. Define $\mathcal{S}_{\mathcal{V},A}$ as the set of tokens, {"Cha cha", "Jive", "Quickstep", "Rumba", "Tango", "Waltz"}. Define the success criteria as: "the amount of ground truth reproduced is inconsistent with random selection." Provided BALLROOM is a sample of $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$, it is relevant to this use case; and depending on its size relative to the variability of the population (which is predicted, e.g., using expert elicitation), a measurement of the ground truth reproduced by a system tested in BALLROOM could then provide reliable evidence of its success.

A different use case is possible. Define Ω, $\mathcal{R}_\Omega$ and $\mathcal{S}_{\mathcal{V},A}$ as above, but define the success criteria as: "the amount of ground truth reproduced is inconsistent with random selection, and *independent of tempo*." Again, provided BALLROOM is a sample of $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$, it is relevant to this use case. However, a measurement of the amount of ground truth reproduced by a system tested in BALLROOM is not relevant to the use case *because* it does not control for the restriction imposed in the success criteria. A different evaluation must be designed.

The use case proposed in [30] is not the only way or the best way to mitigate ambiguity in MIR research, but I claim that it is one way by which a research problem and task can be defined with clarity, and which thereby can aid with the design and evaluation of MIR systems.

3.2 On the formal design of experiments

The aim of DOE is to help one produce the most reliable evidence at the least cost. DOE is an area of statistics that has become essential to progressive science and profitable industry, from biology and genetics to medicine and agriculture [1]. (In fact, it arose from agriculture in the early 20th century.) Hence, I claim that it is reasonable to argue that DOE can help build a formal framework of evaluation that can reliably guide the engineering of systems to address the aims of MIR (Sec. 1.1).

The design of an experiment entails performing several essential and non-trivial steps. In the terminology of DOE [1], this includes identifying treatments, experimental and observational units, identifying structures in the treatments and plots, creating the design, and specifying the response model, all with respect to the hypothesis or quantity under investigation. Below are some definitions of these components. Table 1 provides examples.

Treatments Descriptions of what is applied to an experimental unit, indexed by $\mathcal{T} = \{1, \ldots, t\}$.

Experimental unit The smallest unit to which a treatment is applied.

Observational unit (plot) The smallest unit on which a response is measured. The set of N plots is indexed by $\Omega := \{\omega : \omega \in \{1, \ldots, N\}\}$.

Experimental design A map $T : \Omega \to \mathcal{T}$.

Plot/Treatment structure Meaningful ways of dividing up the plots/treatments.

Response model An assumed mathematical relation between the response and the treatment parameter.

Treatment Parameter The (latent) contribution of the treatment to the measured response.

Fundamental questions that DOE answers are: for my t treatments, how large should N be to reach my required experimental power and not exceed my resources? How should I collect the plots? How should I map the plots to the treatments? An essential part of answering these is "expert elicitation," whereby knowledge about the plots and treatments is collected from someone familiar with them. This informs the design, response model, and subsequent analysis. For example, it is important to know in experiments (a) and (c) in Table 1 if the plots have structure, e.g., positions of pots in greenhouse get variable sunlight; and students in the course are math majors or non-math majors. Otherwise, if the amount of sunlight correlates with the amount of water and compost applied to a plant, or if most students that take the course remotely are math majors, then one might conclude that compost and water have a negative effect on yield (confounding of sunlight and treatment), or remote learning is better than local learning (confounding of student background and learning method).

At first it seems standard MIR tasks and evaluation need only be "translated" into the language of DOE, and then existing statistical machinery be deployed [32]. This translation is not so immediate, however. Consider the evaluation performed in [33]: 100 repetitions of stratified 10-fold CV (10fCV) in a specific dataset sampled from some $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$. In each repetition, several systems are trained and tested, the mean amount of ground truth reproduced by the resulting systems is measured, the mean and standard deviation of the 100 repetitions are reported, and conclusions are made. This appears to be a factorial design, crossing F (feature extraction method) and M (supervised learning method). The treatments then appear to be all levels of $F \wedge M$. The experimental and observational unit then is a complete 10fCV, of which there are $N = 100|F||M|$, i.e., each level in $F \wedge M$ treats 100 10fCV plots. Finally, the response is the proportion of ground truth reproduced.

This scenario appears similar to testing for a significant difference between local and remote learning in Table 1(c), except there are some important differences. First, any pair of systems in each level of $F \wedge M$ in any repetition of 10fCV share 80% of the same training data. Hence, the 10 systems produced in each level of $F \wedge M$ in any repetition of 10fCV are not independent. Second, all repetitions of 10fCV at a level of $F \wedge M$ produce measurements that come from the same data. Hence, the 100 plots are not independent. Third, the systems themselves are generated from training data used to test other systems produced at the same level. Considering the scenario in Table 1(c), this would be like tailoring the implementations of local and remote classes to some of the students in them. This means the realisations of the treatments come from material that they in turn treat, which introduces a non-trivial dependency between treatments and plots. Finally, there is unacknowledged structure in the particular dataset used in [33], which can bias the response significantly [26].

All of the above and more [10] means that the responses measured in this experiment should be modelled in a more complex way than the simple textbook model [1] — which assumes a normal distribution having a variance that decreases with the number of measurements, and a mean that is the treatment parameter (in this case, the expected generalisation of a level in $F \wedge M$ in $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$). What can be concluded from the kind of experiments in [33] remains to be seen, but it is clear that not much weight should be given to conclusions drawn from the simple textbook model [32].

3.3 On the feedback

The initial evaluation of the three systems in Sec. 1.5 is of little use for improving them with respect to the problem they are designed to address. It even provides misleading information about which is best or worst at addressing that problem. The knowledge produced by the evaluation is thus suspicious. Instead, my system analysis, along with the intervention experiment in Sec. 1.6, provide useful knowledge for improving the systems, as well as evaluation using BALLROOM. Our work in [21] provides another example of this. According to the linchpin challenge $\mathrm{R_{II}}$ in Sec. 1.3, and its discussion in the Roadmap, I claim that system analysis along with the intervention experiment — designed to explain why a system reproduces an amount of ground truth inconsistent with random selection — can lead to real and useful knowledge for improving MIR systems and evaluation practices.

[10] There are other major issues that contribute complications, e.g., the meaning of $\mathcal{R}_\Omega \times \mathcal{S}_{\mathcal{V},A}$, whether "ground truth" is a rational concept, and the problem of "curation" in assembling of a music dataset.

4. CONCLUSION

The MIR discipline has reached a level of maturity such that its impact is undeniable [3,5], and its leaders recognise contemporary needs for targeted development in specific directions [23]. In this position paper, I address the direction of improving MIR evaluation — specifically the linchpin challenges $\mathrm{R_I}$ and $\mathrm{R_{II}}$ (Sec. 1.3) — and revisit several priorities for improving evaluation. I argue that these only treat the symptoms of the problem and not its cause: MIR lacks a formal evaluation framework relevant to its aims. I argue that addressing this cause is the principal priority. I propose that this can be addressed by leveraging established DOE along with an effort to mitigate ambiguity in research. However, this is not as straight-forward as I initially envisioned [26, 32]. To develop and integrate such a framework into the MIR research pipeline will require far more effort than is currently being devoted to it. It will require the focus of a multidisciplinary team of specialists for many years. Toward this end, I hope this position paper persuades some to participate in solving the core problem.

4.1 Acknowledgments

This paper comes from feedback to many grant proposals, paper submissions, poster discussions, public seminars, and private conversations. I give considerable thanks to all who have taken the time to talk, and to: H. Maruri-Aguilar, B. Parker, F. Rodríguez-Algarra and H. Grossmann.

5. REFERENCES

[1] R. A. Bailey. *Design of comparative experiments*. Cambridge University Press, 2008.

[2] T. Bertin-Mahieux, D. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proc. ISMIR*, pages 591–596, 2011.

[3] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE*, 96(4):668–696, Apr. 2008.

[4] D. Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3), 11 2014.

[5] S. J. Cunningham, D. Bainbridge, and J. S. Downie. The impact of MIREX on scholarly research. In *Proc. ISMIR*, pages 259–264, 2012.

[6] M. Davies and S. Böck. Evaluating the evaluation measures for beat tracking. In *Proc. ISMIR*, pages 637–642, 2014.

[7] S. Dixon, F. Gouyon, and G. Widmer. Towards characterisation of music via rhythmic patterns. In *Proc. ISMIR*, pages 509–517, 2004.

[8] J. Downie, A. Ehmann, M. Bay, and M. Jones. The music information retrieval evaluation exchange: Some observations and insights. In *Advances in Music Information Retrieval*, pages 93–115. Springer, 2010.

[9] J. S. Downie. The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28(2):12–23, 2004.

[10] A. Flexer. Statistical evaluation of music information retrieval experiments. *J. New Music Research*, 35(2):113–120, 2006.

[11] A. Flexer. On inter-rater agreement in audio music similarity. In *Proc. ISMIR*, pages 245–250, 2014.

[12] D. J. Hand. Deconstructing statistical questions. *J. Royal Statist. Soc. A (Statistics in Society)*, 157(3):317–356, 1994.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2 edition, 2009.

[14] John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2007.

[15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[16] O. Nieto, M. M. Farbood, T. Jehan, and J. P. Bello. Perceptual analysis of the f-measure for evaluating section boundaries in music. In *Proc. ISMIR*, pages 265–270, 2014.

[17] G. Peeters, J. Urbano, and G. J. F. Jones. Notes from the ISMIR 2012 late-breaking session on evaluation in music information retrieval. In *Proc. ISMIR*, 2012.

[18] A. Pikrakis. A deep learning approach to rhythm modeling with applications. In *Proc. Int. Workshop Machine Learning and Music*, 2013.

[19] A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra. Acousticbrainz: A community platform for gathering music information obtained from audio. In *Proc. ISMIR*, pages 786–792, 2015.

[20] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. mir eval: A transparent implementation of common MIR metrics. In *Proc. ISMIR*, pages 367–372, 2014.

[21] F. Rodríguez-Algarra, B. L. Sturm, and H. Maruri-Aguilar. Analysing scattering-based music content analysis systems: Where's the music? In *Proc. ISMIR*, 2016.

[22] M. Schedl, A. Flexer, and J. Urbano. The neglected user in music information retrieval research. *J. Intell. Info. Systems*, 41(3):523–539, 2013.

[23] X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordà, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. *Roadmap for Music Information ReSearch*. Creative Commons, 2013.

[24] B. L. Sturm. Evaluating music emotion recognition: Lessons from music genre recognition? In *Proc. ICME*, 2013.

[25] B. L. Sturm. A simple method to determine if a music information retrieval system is a "horse". *IEEE Trans. Multimedia*, 16(6):1636–1644, 2014.

[26] B. L. Sturm. The state of the art ten years after a state of the art: Future research in music information retrieval. *J. New Music Research*, 43(2):147–172, 2014.

[27] B. L. Sturm. A survey of evaluation in music genre recognition. In A. Nürnberger, S. Stober, B. Larsen, and M. Detyniecki, editors, *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, volume LNCS 8382, pages 29–66, Oct. 2014.

[28] B. L. Sturm. Faults in the latin music database and with its use. In *Proc. ISMIR (Late breaking demo)*, 2015.

[29] B. L. Sturm. The "horse" inside: Seeking causes of the behaviours of music content analysis systems. *ACM Computers in Entertainment (accepted)*, 2015.

[30] B. L. Sturm, R. Bardeli, T. Langlois, and V. Emiya. Formalizing the problem of music description. In *Proc. ISMIR*, pages 89–94, 2014.

[31] B. L. Sturm, C. Kereliuk, and A. Pikrakis. A closer look at deep learning neural networks with low-level spectral periodicity features. In *Proc. Int. Workshop on Cognitive Info. Process.*, pages 1–6, 2014.

[32] B. L. Sturm, H. Maruri-Aguilar, B. Parker, and H. Grossmann. The scientific evaluation of music content analysis systems: Valid empirical foundations for future real-world impact. In *Proc. ICML Machine Learning for Music Discovery Workshop*, 2015.

[33] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302, July 2002.

[34] J. Urbano, B. McFee, J. S. Downie, and M. Schedl. How significant is statistically significant? the case of audio music similarity and retrieval. In *Proc. ISMIR*, pages 181–186, 2012.

[35] J. Urbano, M. Schedl, and X. Serra. Evaluation in music information retrieval. *J. Intell. Info. Systems*, 41(3):345–369, Dec. 2013.

[36] K. West, A. Kumar, A. Shirk, G. Zhu, J. S. Downie, A. Ehmann, and M. Bay. The networked environment for music analysis (nema). In *World Congress on Services*, pages 314–317, July 2010.